Research Report

A Scale Drift Study

Jinghua Liu

Edward Curley

Albert Low

December 2009

ETS RR-09-43

A Scale Drift Study

Jinghua Liu and Edward Curley

ETS, Princeton, New Jersey

Albert Low

National Board of Medical Examiners, Philadelphia

December 2009

Abstract

This study examines the stability of the SAT® scale from 1994 to 2001. A 1994 form and a 2001 form were readministered in a 2005 SAT administration, and the 1994 form was equated to the 2001 form. The new conversion was compared to the old conversion. Both the verbal and math sections exhibit a similar degree of scale drift, but in opposite directions: the verbal scale has drifted upward, whereas the math scale has drifted downward. We suggest testing programs monitor the score scales periodically by building a testing form schedule that allows a systematic and periodic checking of scale stability.

Key words: Scale drift, scale stability, SAT, monitoring scale

For testing programs that continuously administer multiple forms across years, score equating (Kolen & Brennan, 2004) is used to adjust for form difficulty differences so that scores are comparable and can be used interchangeably. Scores are usually reported on a score scale, into which test developers incorporate meaning to facilitate the interpretation of scores by test users. For example, the SAT scores are reported on the College Board 200-to-800 scale. A scaled score of 700 on the SAT math test indicates the same ability level regardless of which form is administered to a test-taker.

However, equating is not a panacea for assuring score comparability. Scaled scores for testing programs can become less comparable over time. For instance, the SAT scale was recentered in 1995. The most salient reason for this recentering was that the expected mean verbal and math scores when the scale was originally set up during the 1940s had become considerably different for groups who took the test five decades later (Dorans, 2002). A score of 700 on the SAT math test in 1995 would have meant something different than a score of 700 in 1940 if the recentering had not been done. Scale drift occurred here.

Haberman and Dorans (2009) defined scale drift as when there is a shift in the meaning of a score scale that alters the interpretation that can be attached to score points along the scale. Scale drift may occur due to a variety of reasons. For instance, the norms group used at the time the scale was established may not be appropriate over time, as mentioned above in the SAT recentering case. Also, equating is imperfect both due to violations of equating assumptions and due to use of finite samples to estimate parameters (Haberman, Guo, Liu, & Dorans, 2008). Concerns about equating assumptions being violated may stem from evolution of test content, curricula or populations over time or from violations of invariance assumptions. Other imperfections involve accumulated error in equating models when that error is consistently in the same direction (Livingston, 2004). “Even though an equating process can maintain the score scale for some time, the cumulative effects of changes might result in scores at one time being not comparable with scores at a later time” (Kolen, 2006, p. 169).

These concerns lead to professional standards recommending that evidence be compiled periodically to document the stability of a score scale, thereby ensuring appropriate use of test scores over significant periods of time (ETS, 2002). According to Standard 4.17, “Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 59).

The purpose of this study is to identify whether there has been scale drift and to determine the extent of scale drift, if there is any, on the SAT Reasoning Test™. The SAT measures critical thinking skills that students have developed over time and that they need in order to be academically successful in college. Each year, there are seven SAT administrations and more than 2 million students take the SAT. Nearly every college in America uses the test as a common and objective scale for evaluating a student’s college readiness. It is a high-stakes test, and it is important to maintain consistency of the scale to ensure the score comparability over time. We hope that what we report can shed light on this research and practical area, and that testing programs can make an effort to incorporate the examination of scale drift into operational routine procedures.

Previous Research on Scale Drift

The SAT program has been periodically monitoring the scale. Stewart (1966) examined the extent of drift that might have occurred in the SAT-verbal (SAT-V) scale since 1941. Four old forms from 1944, 1948, 1953 and 1957, respectively, were equated to a 1963 form. The newly derived scaled scores of each old form were compared with the original scaled scores when each old form was initially equated. The results showed that: (a) scores reported in 1963 might not have the same meaning as scores reported in 1944, (b) SAT-V scores might have been 20 to 35 points higher in 1963 than in 1948, and (c) for both the 1953 and 1957 forms, the scaled score differences were less than 5 points, warranting the conclusion that the SAT-V scale remained relatively stable during the period from 1953 through 1963.

Modu and Stern (1975) assessed the SAT-V and SAT-math (SAT-M) scales between 1963 and 1973. Old forms from 1963 and 1966 were equated to a 1973 form. The results indicated that the conversions derived in 1973 for the old forms were higher than the 1963 and 1966 conversions, in both the verbal and math sections. The average difference was 14 points for verbal and 17 points for math scores. Both studies employed a nonequivalent-group anchor test design. An anchor test from an old form was embedded in a new form administration. Then the old form was equated to the new form through an anchor.

The last time a scale drift study was conducted for the SAT was in 1994 (McHale & Ninneman, 1994). The stability of the scale was assessed by placing the 1973-1974 forms and the 1974-1975 forms on the scale of 1983-1984 forms. Two equating designs were implemented. In the first design, a nonequivalent-group anchor test (NEAT) design was used. Anchors from three 1973-1974 forms were embedded in the 1983-1984 administrations. In the second design, the operational sections of two 1974 and 1975 SAT forms were readministered at 1984 administrations. The 1984 booklets contained one of the operational sections (V1, V2, M1 or M2) of the 1974 forms. Section preequating was conducted to obtain the new conversion for the 1974 forms. The results indicated that the SAT-V scale was relatively stable from 1973 to 1984, with little or no drift. However, the math results were inconsistent. The external anchor equatings indicated that the math scale had drifted upward an average of 6 to 13 points, whereas the equatings based on readministration of the old SAT forms indicated that the math scale had drifted downward an average of 6 to 14 points.

Scale drift issues have also been explored from other perspectives. Peterson, Cook and Stocking (1983) examined several equating methods in terms of how robust those methods are in maintaining scale stability. The results indicated that linear equating methods perform adequately when the tests are reasonably parallel. When the tests differ in content and length, methods based on the IRT model lead to greater stability of equating results.

Puhan (2009) examined the scale drift for a testing program that employs a cut score. He compared the scaled scores that were on the same form, but derived from different lengths of equating chains (i.e., the new form was equated to the base form through two intermediate forms in Chain 1, whereas the same new form was equated to the same base form through five different intermediate forms in Chain 2). Results indicated that there were some differences between the conversions derived via different equating chains.

Overall, the previous studies all revealed that scale drift can occur. The current SAT score scale was recentered in 1995 and scores have been reported on the new scale since then (Dorans, 2002). It is critical that the scale stability be reassessed to ensure the consistency of score meaning.

Methodology

Design

In a typical scale drift study, an old form is spiraled, either in intact form or in sections, along with a new form. The old form is equated to the new form to obtain a new raw-to-scale conversion. This new raw-to-scale conversion is then compared to the original raw-to-scale conversion of the old form. Large differences suggest instability of the scale.

In this study, we employed such a typical design: We identified an old form, spiraled it with a newer form and administered both forms in the same SAT administration. It was not easy, though, to find such an old form and a newer form due to several reasons. First, since the SAT was redesigned in 1994, it precluded the use of any forms prior to 1994. Second, when this study was designed in 2004, a newly redesigned SAT was going to be launched in March 2005, which precluded the use of any forms after March 2005. Also, since the test was being redesigned, fewer new forms were developed than usual and reprints were used more often. We would have liked to choose two forms that were as distant from one another as possible, but the pool of available forms was limited. With all the constraints, we found a 1994 form and a 2001 form that could be administered at the January 2005 administration. Note that the original conversion for the 1994 form was transformed to a conversion on the recentered SAT scale so that the original conversion could be compared to the new conversion.

Although this is a typical design for collecting data for a scale drift study, parts of it are nonstandard. As can be seen from Figure 1, the standard process for assessing scale drift between 1994 and 2005, for example, would be to readminister a 1994 form in 2005, along with a new 2005 form, and equate the 1994 form to the 2005 form. The equating results for the 1994 form equated to the 2005 form in 2005 can then be compared to the original 1994 conversion. Any differences between the two conversions can be attributed to scale drift occurring between 1994 and 2005.

What is different in our study is that there was no available form like the 2005 form discussed above. Instead, the 1994 form was readministered with a 2001 form in the 2005 administration (see Figure 2). The 1994 form was equated to the 2001 form, based on data from the 2005 administration. What makes our design even more nonstandard is the form revision that took place. Both forms were reviewed by ETS content experts to ensure that none of the questions has become outdated or inaccurate. As a result, no changes were required to the 2001 form, but three math questions were replaced in the 1994 form because of changes between 1994

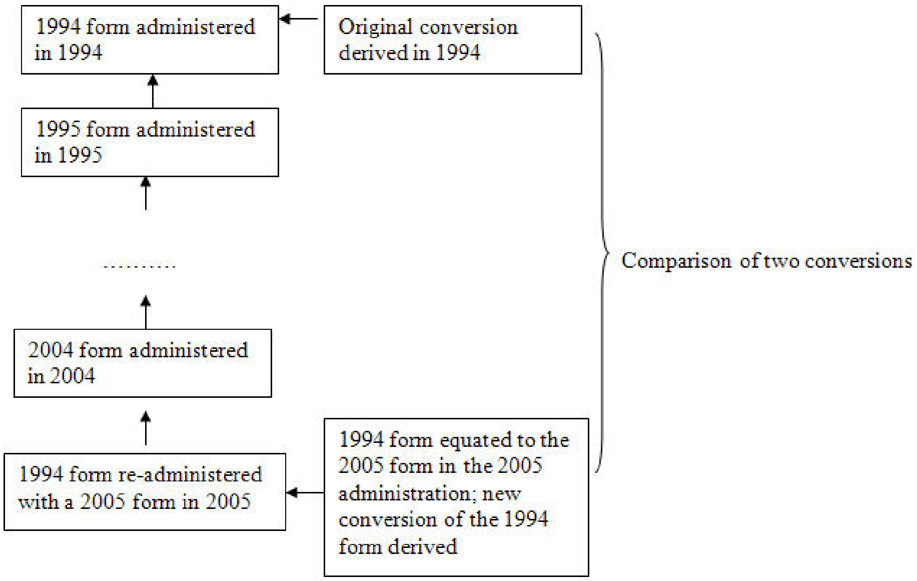
1994 form administered in 1994
Original conversion derived in 1994
1995 form administered in 1995
....
2004 form administered in 2004
1994 form re-administered with a 2005 form in 2005
1994 form equated to the 2005 form in the 2005 administration; new conversion of the 1994 form derived
Comparison of two conversions

Figure 1. **The standard process of a scale study design.**

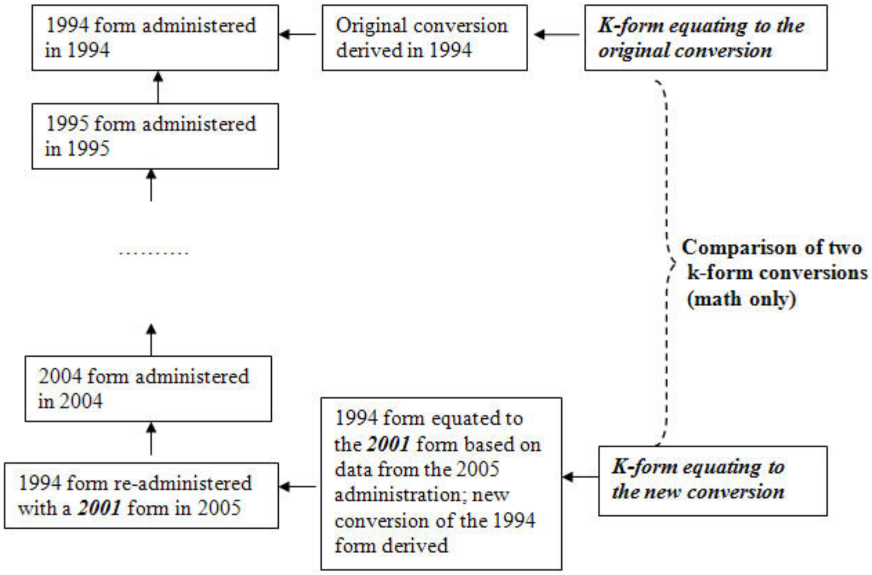
1994 form administered in 1994
Original conversion derived in 1994
K-form equating to the original conversion
1995 form administered in 1995
....
Comparison of two k-form conversions (math only)
2004 form administered in 2004
1994 form re-administered with a 2001 form in 2005
1994 form equated to the 2001 form based on data from the 2005 administration; new conversion of the 1994 form derived
K-form equating to the new conversion

Figure 2. **The nonstandard process of the scale study design in the current study.**

and 2005 in the particular calculators allowed for use during SAT administrations. Hence, only 57 of the 60 math questions in this form were identical in the 1994 and 2005 administrations. Therefore, the conversions based on 60 items could not be directly compared. A k-form equating was conducted on the 1994 form to equate the 57-item test to the 60-item test based on data from the 1994 administration, and then a k-form equating was conducted on the 1994 form to equate the 57-item test to the 60-item test based on data from the 2005 administration. The two k-form conversions based on 57 items were then compared. Figure 2 displays the nonstandard procedure employed in this study. The nonstandard parts are highlighted in bold and italics in Figure 2.

Test Forms

Both the 1994 form and the 2001 form had three verbal sections, three math sections and a variable section. In total, there were 78 verbal items and 60 math items that were counted toward the reported scores for verbal and math, respectively.

Equating Design and Equating Methods

In a normal SAT administration where a new form or forms are administered, there are two types of data collection designs employed for score equating: the nonequivalent groups anchor test (NEAT) design and the equivalent groups (EG) design. At each SAT administration where one new form is administered, the new form is equated to four old forms through a NEAT design. An EG design is usually employed in an SAT administration where two new forms are administered, the first new form is equated to four old forms using the NEAT design and the second new form is equated to the first new form through an EG design

The 1994 form is equated to the 2001 form at the 2005 administration. In this study, we did not use the NEAT design. Instead, we used the original raw-to-scale conversion for the 2001 form, and then we equated the 1994 form to the 2001 form via an EG design.

The 1994 form was spiraled with the 2001 form and both forms were administered at the 2005 administration. Spiraling refers to packaging and distributing two or more test forms in an alternative sequence (e.g., Form1, Form 2, Form 1, Form 2, and so on). The spiraling procedure used in the SAT administration and the large number of test-takers who take each form usually ensure equivalent groups at the same administration. The 1994 form was equated to the 2001 form at the 2005 administration via an EG design. Presmoothing was performed for both forms using a loglinear univariate model, preserving six marginal moments (Holland &

Thayer, 2000). Both linear equating (mean-sigma) and direct equipercentile equating were conducted. The equipercentile equating was deemed the most appropriate for both the verbal scores and the math scores.

The 1994 k-form equatings for math. A k-form is created when slight revisions are made to an original form. Such revisions may include dropping or replacing a few items, revising items based on changed wording, or from adding additional information to an item. A k-form equating is a process that links the slightly revised form, the k-form, back to the original form via a single group design.

For the math section, since three items were replaced in the 1994 form when given in 2005, the 60-item conversion resulting from equating the 1994 form to the 2001 form based on data from the 2005 administration could not be directly compared to the original 60-item conversion based on data from the 1994 administration. Instead, a 57-item k-form equating was conducted using data from the 1994 administration; and another 57-item k-form equating was conducted using data from the 2005 administration. The two k-form conversions were compared to each other. See Figure 2. Again, linear equating (mean-sigma) and direct equipercentile equating were conducted. The equipercentile equating was deemed the most appropriate for the two k-form equatings.

Discrepancy Indices

Both the differences of unrounded scaled scores between the original conversion and the newly derived conversion for verbal and the differences between the two k-form conversions for math were evaluated graphically and analytically.

Difference plots of conversions. The difference plot, the new conversion minus the old conversion, is the most direct means of assessing score differences. At each score level, the new conversion is compared to the old conversion and the difference in the conversions is plotted.

To evaluate the relative magnitude of a difference in score conversions, we use the differences that will have practical consequences as a criterion, which we call minimum differences that might matter (MDTMM). On the SAT scale, scores are reported in 10-point units (i.e., 200, 210, 220…780, 790, 800). If the two unrounded conversions differ by no more than 5 points, then ideally the scores should be rounded to the same scaled score. However, this does not always happen due to rounding. For example, at a raw score of 53, the corresponding unrounded scaled score might be 784.9, based on the 1994 conversion, and 785.1, based on the new conversion derived in 2005. Due to rounding, the rounded reported scores would be 780 for the 1994 conversion, and 790 for the new conversion derived in 2005, when ideally they should be identical. The MDTMM, in contrast, treats these two conversions as being equivalent at this raw score point. Dorans, Holland, Thayer, and Tateneni (2003) adapted this notion to other tests and considered MDTMM to be half of a score unit for unrounded scores. In the present study, we use half of the SAT score unit, 5 points, as the criterion. Note this difference is best thought of as an indifference threshold. Any differences less than the MDTMM are considered not big enough to warrant any concern since they are smaller than the smallest difference that might actually matter.

Root expected square difference (RESD). We calculated the root expected square difference (RESD) statistic, which is given by the formula

$$RESD = \sqrt{\frac{1}{N}\sum f_x[s_{new}(x) - s_{old}(x)]^2} \,, \tag{1}$$

to provide a standardized measure that evaluates the extent of the difference between the new conversion s_{new} and the original conversion s_{old}, where *N* is the total number of test-takers taking the new form, and f_x is the frequency at each score point. Note that we expressed the differences in scaled score (*S*) units rather than in the raw score units, because most readers can understand and readily interpret the differences on the familiar College Board 200-to-800 scale.

Percentage of score/examinees exceeding MDTMM. In addition to using RESD, we made use of the percentage of raw scores for which the new conversion differed from the old conversion by more than five points, and the percentage of examinees for whom these conversions create differences of more than five points. These two indices provide straightforward insight into lack of stability as a percentage of the score range, and as a percentage of the test-takers. The calculation of the two percentage indices is

$$D_x = 1 \; if \; | S_{new}(x) - S_{old}(x) | >= MDTMM$$

$$\% Affected \; Score \; Points = \frac{\sum_x D_x}{X_{\max} - X_{\min} + 1}$$

$$\% Affected \; Examinee = \frac{\sum_x f_x D_x}{N}. \qquad (2)$$

Results

The results presented below were obtained using equipercentile equating. The original raw-to-scale conversion was compared with the newly derived raw-to-scale conversion. The MDTMM criterion was used to evaluate the differences.

Table 1 provides descriptive statistics for verbal and math scores on the 1994 form and on the 2001 form when given in the 2005 administration. Data in Table 1 show that the verbal mean of the 1994 form was slightly lower than that of the 2001 form, indicating that the 1994 verbal section was more difficult than the 2001 verbal section since the groups taking each form were equivalent. On the other hand, the math mean of the 1994 form was slightly higher than that of the 2001 form, suggesting that the math section of the 1994 form was slightly easier than the math section of the 2001 form.

Table 1

Descriptive Statistics for the 1994 Form and the 2001 Form in the 2005 Administration

	Verbal		Math	
	1994 form	2001 form	1994 form	2001 form
N	89,680	91,733	89,680	91,733
Mean	32.99	33.56	27.03	26.52
SD	16.80	17.29	15.06	14.37
Skewness	0.21	0.26	0.16	0.09
Kurtosis	2.35	2.33	2.14	2.21

Verbal Section

Figure 3 presents the differences of unrounded scaled scores between the original 1994 conversion and the newly derived conversion when the 1994 form was equated to the 2001 form based on data from the 2005 administration (new – old). When the dashed curve is above the scaled score difference of zero, it represents an increase from the original conversion to the new conversion. On the other hand, as the curve drops below the scaled score difference of zero line, it represents a decrease from the original conversion to the new conversion. As can be seen in Figure 3, the new conversion was higher than the old conversion across most of the score range. In other words, the 2005 test takers would have obtained slightly lower verbal scores if the original 1994 conversion had been applied. Further, the difference curve fell within the MDTMM range (5 scaled score points) below scaled score of 300 and between score of 500-800. However, the differences were larger than 5 in the range of 300-500. These differences were nonnegligible.

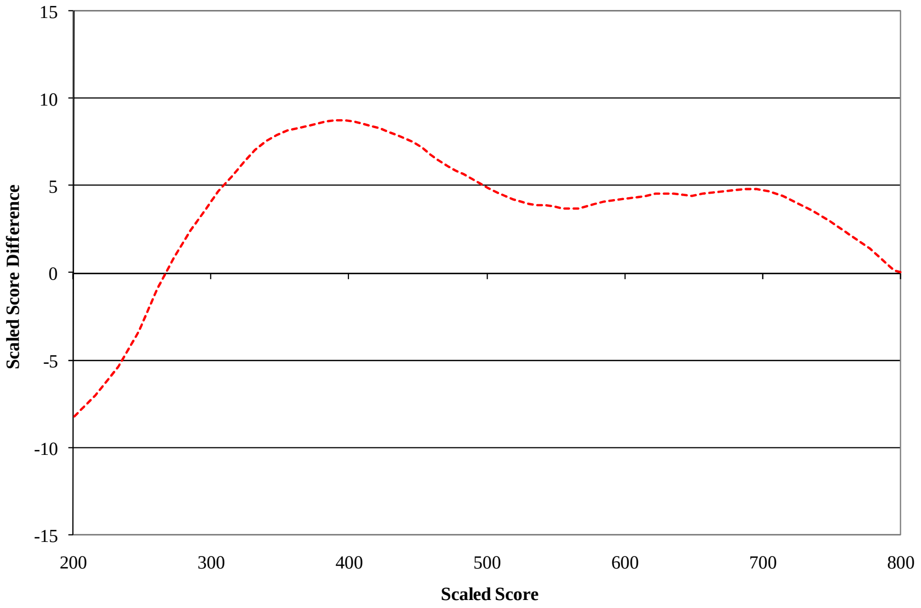

***Figure 3*. The raw-to-scale differences between the 1994 verbal conversion and the verbal conversion when the 1994 form was equated to the 2001 form in the 2005 administration.**

Table 2 summarizes the differences between the two conversions. For the 2005 test takers, means and standard deviations based on the original conversion and based on the new conversion are listed respectively. Also listed are the difference in means, the RESD value, the percentage of raw scores with an absolute unrounded scaled score difference equal to or larger than 5, and the percentage of examinees whose conversions resulted in scores that differ by at least 5 points. The mean based on the new conversion was 487, 7 points higher than the mean based on the original conversion. The RESD value is 6, which is larger than the MDTMM (5), suggesting a nonnegligible difference. The proportion of raw scores for which scaled scores between the two conversions differed more than 5 points was 31%. The percentage of examinees that would have been affected was 53%. In summary, the results indicate an average drift upward of 6 scaled score points on the verbal section between the new 1994 conversion derived from equating the 1994 form to the 2001 form based on 2005 administration data and the original 1994 conversion.

Table 2

Summary Statistics of Scaled Scores for the 1994 Form Based on the Original 1994 Conversion and Based on the New Conversion Derived via Equating to the 2001 Form Using Data From the 2005 Administration

	Verbal	Math
Sample size	89,680	89,680
Mean & *SD* based on the new conversion derived by equating to the 2001 form using data from the 2005 administration	487 108	496 111
Mean & *SD* based on 1994 conversion	480 110	502 112
Mean difference (new – old)	7	-6
RESD	6	6
% RS with │unrounded scaled score diff│≥ 5	31	55
% Examinees with │unrounded scaled score diff│≥ 5	53	73

Note. RESD = root expected square difference, RS = raw score.

Math Section

Figure 4 displays the differences of two k-form math conversions, based on 57 common items, in the 1994 administration and in the 2005 administration when the 1994 form was equated to the 2001 form. As illustrated in Figure 4, the new conversion was lower than the original 1994 conversion across the entire score range. In other words, the 2005 test takers would have obtained somewhat higher math scores if the original 1994 conversion had been applied. The differences were larger than MDTMM from 200 to 600, and fell within the MDTMM range from 600 to 800.

The differences are summarized in Table 2 as well. The mean based on the newly derived k-form conversion through equating the 1994 form to the 2001 form in the 2005 administration

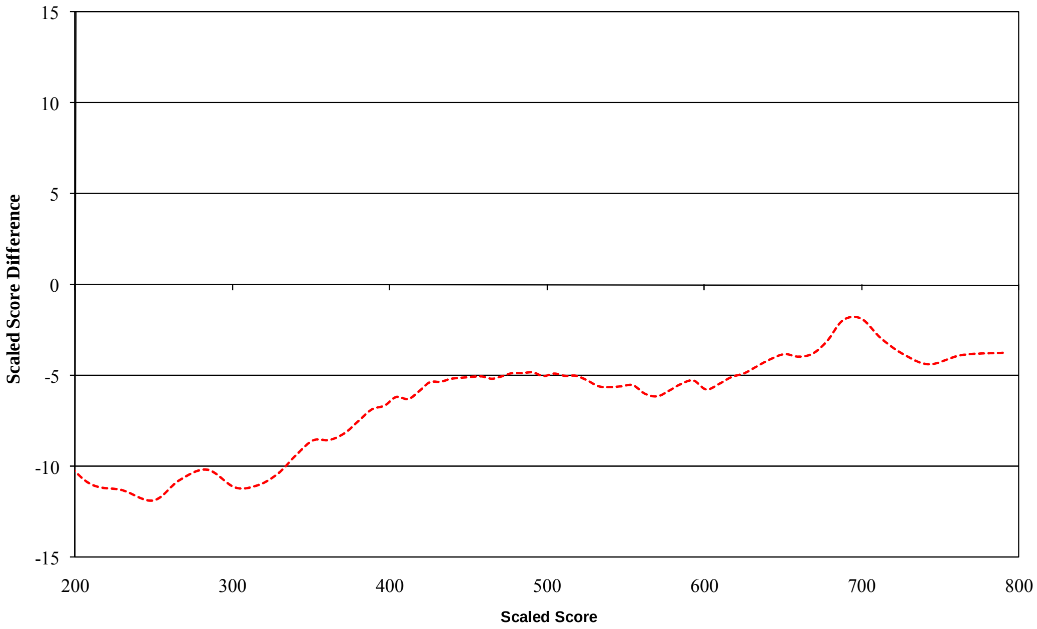

***Figure 4*. The raw-to-scale differences between the 1994 math conversion and the math conversion when the 1994 form was equated to the 2001 form in the 2005 administration.**

was 496, 6 points lower than the mean based on the original 1994 k-form conversion. The RESD value of approximately 6 was larger than the criterion of 5. The percentage of raw scores for which scaled scores between the two conversions differed more than 5 points was 55%. The percentage of examinees for whom the conversions differed by 5 or more was 73%, exhibiting a high degree of departure from the original conversion. In summary, the results indicated that the math scale has drifted downward 6 points on average between the new 1994 conversion derived from equating the 1994 form to the 2001 form based on 2005 administration data and the original 1994 conversion.

An interesting phenomenon observed is that although verbal and math exhibit a similar degree of scale drift, the directions were opposite: the verbal scale has drifted upward, whereas the math scale has drifted downward. In other words, the verbal section appeared more difficult at the 2005 administration than it appeared at the 1994 administration (resulting in a higher raw-to-scale conversion), while the math section appeared easier at the 2005 administration (resulting in a lower raw-to-scale conversion). The score distributions diverged for verbal and math. A possible cause is that test content and/or test takers' familiarity with the content and item types may have shifted. In 1994, the SAT was revised and some major changes took place in both the verbal and math sections. For verbal, the antonyms were removed, and the percentage of questions associated with passage-based reading was increased. For math, two major changes took place: the introduction of student-produce-response (SPR) items, and the use of calculators. The SPR items have become familiar to the test-takers during the 10 years since their introduction. As a result, the test takers in 2005 might perform better than the 1994 test takers on SPR items relative to other items (i.e., passage-based reading items). In addition, test takers in 2005 may be more familiar with using calculators on the test.

It is also interesting that the degree of drift seemed conditioned on the scale range. For verbal and math, the scaled score differences were less than 5 (the MDTMM criterion) at the upper scale range, say above 500. Below 500, the differences were larger. This finding contradicts the Haberman et al. (2008) results, where they found for very high and very low scores, stability results were not as strong as they were for less extreme scores. This issue needs to be explored further.

Discussion

This study examines the stability of the SAT scale from 1994 to 2001. A 1994 form was readministered in a 2005 SAT administration, and was equated to a 2001 form that was also readministered in the same 2005 administration. The new 1994 conversion was then compared to the original 1994 conversion. Graphical and analytical approaches were used to evaluate the differences between the original conversion derived in 1994 and the newly derived conversion based on data from the 2005 administration. The results showed that for both verbal and math, the differences were nonnegligible, suggesting that scale drift might have occurred on both measures.

There are a variety of reasons that scale drift can occur. Haberman and Dorans (2009) categorized some possible sources of scale drift by looking at systematic and random errors. The first source of error identified is population shift. Haberman and Dorans (2009) classified it as a source of systematic error but not a source resulting in scale drift, whereas Petersen (2009) argued that population shift could induce scale drift when there is a mismatch in the ability of the group taking the test and the difficulty of the test, and/or when the population taking the test changes in some way related to performance on the test (e.g., an ESL exam that was initially taken by Asian test-takers and then the test taker population shifts to primarily Latino test-takers). The authors of this report think that population shift could be a source of scale drift. The current SAT scale was based on a 1990 Reference Group. This norms group may not be appropriate after more than 10 years. For example, the mean was set at 500 for both verbal and math when the SAT scale was recentered in 1995. Over the years, both verbal and math means have been progressively increasing. The means based on the 2005 College-Bound Seniors cohort are 508 and 520 for verbal and math, respectively.

Haberman and Dorans (2009) also listed test construction practices as another category of error. In this category, too much variation in the raw-to-scale conversions (i.e., at the same raw score level, a scaled score is 600 for form *X* but 650 for form *Y*), due to loose test-construction practices or due to vague specifications, may lead to scale drift. Construct shift may lead to scale drift; test reliability shift may also lead to scale drift. Even though the SAT forms are assembled based on the very strict blueprint that not only specifies the mean and standard deviation of the test item difficulties, but also calls for a specific number of items at each difficulty level, it is unavoidable that certain variations may have occurred.

The third source of error discussed by Haberman and Dorans (2009) is sampling error. Random error is introduced to equating when finite samples are used to estimate parameters. Accumulation of random errors can cause scale drift when it is consistently in the same direction. Further, using nonrandom, nonrepresentative sampling of examinees can induce scale drift as well.

Another category of error is the use of inadequate anchors or violations of equating assumptions. Factors such as groups being too far apart in ability, the anchors not having a strong correlation with the total tests, or the anchor possessing different content than the total test can result in violations of the equating assumptions and therefore induce scale drift. SAT employs NEAT equating to put the new forms on the scale most of the time. Even though equating to multiple old forms via multiple anchors minimizes certain possible undesirable factors such as those resulting from equating to a single old form, the accumulation of equating errors can be nonnegligible over the time.

What shall we do if scale drift is detected? Here are some recommended procedures:

1. Acknowledge the problem. This is an essential first step to bring the program in compliance with professional standards.

2. Scrutinize test blueprints. Are the blueprints specific enough or too vague? For example, do the statistical specifications only specify a mean and standard deviation of the item difficulty for the test, or specify the number of items at different difficulty levels as well? Or is a target information curve used? Vague specifications open the door for scale drift to sneak in. It may be necessary to consider strengthening vague specifications to make them more specific.

3. Examine the test assembly process. When the test is assembled, how strictly are the blueprints followed? Is there too much leeway exercised during the assembly process?

4. Examine the construct being measured. Has the content been changed? Have the measurement conditions changed? Do certain types of items favor or disfavor certain groups, even though DIF at the item level is small?

5. Investigate test administration conditions. Are the tests administered under the same conditions, in terms of testing modes, testing timing, and so on? Did

anything unusual happen during a specific administration in certain test taking regions to affect the equating results?

6. Scrutinize statistical analysis processes. How are the data collected? How is the equating done? Are any of the equating assumptions seriously violated? Are the equating samples large and representative? Is the choice of anchor test reasonable? Are the two populations taking the new form and the old form similar enough to ensure sound equatings?

For future research and operational practice as well, we highly recommend that testing programs monitor the score scales systematically and periodically. For example, build a form schedule that allows the readministration of an old form along with a new form every 5 years. Equate the old form to the new form and then compare the new conversion with the original conversion. A 5-year interval is a good time frame, in that the content of the old form usually does not become outdated in such a time period. We also recommend using an equivalent groups design rather than using the NEAT design (meaning readminister the entire old form rather than readministering part of the old form that can be used as an anchor to link the old test to the new test). The EG design is superior to the NEAT design because the differences in performance are due to differences in tests, not in tests and groups.

However, this study examined only one old form. Administration of an old form has both statistical and substantive problems. The statistical issue involves the limited data from use of a single form and whether the conclusions can be generalized. The substantive issue involves the suitability of an old form for administration given changes in curricula, in test specifications and in test-taking population (even within the time frame of 5 years, we still need to face these problems). In such a case, differences in equating results may be difficult to interpret. Given all that, it is not surprising that the finding from this study is not quite consistent with the study by Haberman et al. (2008), where they examined time series composed of mean scaled scores and raw-to-scale conversions for 54 SAT verbal and math forms administered from April 1995 to December 2003. They found that the data provide a picture of stability on the whole. For future research, the exploration of designs and methods that do not use old forms to assess scale drift is something worth exploring.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Dorans, N. J. (2002*). The recentering of SAT scales and its effect on score distributions and score interpretations* (College Board Report No. 2002-11). New York: College Entrance Examination Board.

Dorans, N. J., Holland, P.W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS Research Rep. No. RR-03-27, pp. 79-118). Princeton, NJ: ETS.

ETS. (2002*). ETS standards for quality and fairness*. Princeton, NJ: Author.

Haberman, S., & Dorans, N. J. (2009, April). Scale consistency, drift, stability: Definitions, distinctions and principles. In J. Liu & S. Haberman (Chairs), *Inconsistency of scaling function: Scale drift or sound equating?* Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Haberman, S., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning Test score conversions* (ETS Research Rep. No. RR-08-67). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133-183.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Livingston, S. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

McHale, F. J., & Ninneman, A. M. (1994). *The stability of the score scale for the scholastic aptitude test from 1973 to 1984* (ETS Statistical Rep. No. SR-94-27). Princeton, NJ: ETS.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (ETS Research Bulletin No. RB-75-9). Princeton, NJ: ETS.

Petersen, N. S. (2009, April). Inconsistency of scaling function: Scale drift or sound equating? Discussion. In J. Liu & S. Haberman (Chairs), *Inconsistency of scaling function: Scale drift or sound equating?* Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education, 22*, 79-103.

Stewart, E. E. (1966). *The stability of the SAT-Verbal Score Scale* (ETS Research Bulletin No. RB-66-37). Princeton, NJ: ETS.